Fields Model Initiative Interpretability Challenge

Fields Model Initiative Team[1,2]
FMI Interpretability Competition Organizers[1,2,3,4,5,6]

[1] LLMC, National Institute of Informatics, Japan
[2] University of Oxford
[3] University of Tübingen
[4] Sapienza University of Rome
[5] Munich Center for Machine Learning
[6] Transformers Club, Masaryk University

aimo.interp@gmail.com

Abstract. This paper presents a call for participation for interpretability researchers interested in reasoning models, newly opened by AIMO[7] within the Fields Model Initiative (FMI). With 10 million dollars in prizes and a broad media coverage, last year, AIMO attracted the submission of over 2,000 AI models, with the top ones achieving a state-of-the-art in mathematical reasoning. This year[8], FMI sets off with a mission to turn this enormous engineering effort into a better understanding of reasoning AI. To enable everyone to contribute to the FMI's mission, AIMO will now require all submissions, including implementations, models, and training recipes, to become *fully open*. Additionally, the FMI will support interested researchers with hundreds of H200 GPUs, opening the opportunities of big tech labs to the broader AI and interpretability community. Finally, as part of the FMI, we will organize the FMI Eval and Interp Competition that moves to *quantify* the quality of interpretability methods by challenging their ability to distinguish and *identify* robust reasoning in LLMs.

1 Problem

Superhuman performance of large language models (LLMs) on a broad range of university-grade benchmarks come hand in hand with doubts about the authenticity of their claimed capabilities. This obvious disparity is fueled by the vagueness of AI benchmarks, which rarely show more than a single number.

In the ongoing discussion of the AI community around true LLMs' true capabilities, interpretability research (Interp) can provide an alternative, allowing us to move forward from the circle of *"is it a memorization or not"* by acquiring new, *actionable* knowledge. Previous Interp work shows examples of such work; For instance, it showed that LLMs internally follow representations agnostic to a language [7], explaining the underwhelming results in architectural language

[7] https://aimoprize.com
[8] https://www.kaggle.com/competitions/ai-mathematical-olympiad-progress-prize-3

modularisation [5]. Other work demonstrated that LLMs learn [4] and maintain [8] almost perfectly accurate representations of input numbers, steering the focus of the AI community away from fixing the commonly-assumed, yet non-existent problem. A similar story stems from the functional interpretation of the attention sinking tokens [6] that numerous prior works [1, 3, 2] treated as an unwanted artefact that should be mitigated. Notice how such Interp insights provide much more solid grounds for future work than coarse-grained benchmark numbers.

Nevertheless, such actionability is not standard in interpretability research. Much of the Interp focuses on *qualitative* analyses, which may *not* generalise into a universal knowledge – and thus truly support future work towards better models. Moreover, some of these methods, such as sparse autoencoders, requires exceptional computational resources, making the cutting-edge Interp possible exclusively in big-tech labs, often coupled with their own incentives.

We believe that addressing these challenges will empower Interp to deliver actionable knowledge that can unlock the development of better, more robust LLMs. The ambition of the presented challenge (§3) and competition (§4) is to address these gaps by:

1. Opening up cutting-edge Interp methods to all members of the AI community by providing a scalable compute and open access to state-of-the-art mathematical reasoning models;
2. Turning observations into generalised knowledge by gradually turning Interp into a *quantitative* science — able to measure, compare and, ultimately, incentivize reliability of the findings;
3. Achieving real-world impact by facilitating a dialogue between evaluators and SoTA models creators, while providing sizeable resources necessary for controlled experiments for LLMs' pre- and post-training.

2 AIMO and The Fields Model Initiative

The Artificial Intelligence Mathematical Olympiad competition (AIMO) is an annual competition of AI systems and their abilities to tackle the frontier, olympiad-level problems in mathematical reasoning. Every year, the mathematical community across British universities curates a set of *unseen* reasoning problems (see example in Table 1) in a methodology consistent to the curation of university-level mathematical olympiads. Last year, over 2,000 teams and their AI systems participated in AIMO, including teams from Nvidia, Mistral AI and HuggingFace and achieved extensive media coverage, for instance, in The Wall Street Journal[9] and Bloomberg[10].

This year, The Fields Model Initiative (FMI) will further extend AIMO's impact with a mission to turn the frontier engineering contributions of AIMO teams into new scientific knowledge, accelerating the development of open and

[9] https://www.wsj.com/tech/ai/ai-math-prize-imo-09a5852d

[10] https://www.bloomberg.com/news/articles/2024-12-05/billionaire-gerko-s-xtx-gives-millions-to-make-ai-better-at-math

Problem	Result
A 500×500 square is divided into k rectangles, each having integer side lengths. Given that no two of these rectangles have the same perimeter, the largest possible value of k is K. What is the remainder when K is divided by 105?	520

Table 1. Example of AIMO problem.

reliable LLMs for mathematics. This will be achieved by a close engagement with the best-performing AIMO teams in the fully open reproduction of their training methods and a design of scientific ablations rigorously documenting the covariates of their success. The FMI effort will be supported by the AIMO organizing team and the National Institute of Informatics in Japan, committed to provide the AIMO teams with (1) an assistance of researchers and (2) a computing power of hudreds of H200 GPUs.

FMI is planned as a follow-up phase of the AIMO competition and will be open for participants between December 2025 and June 2026, with AIMO's top-ranked reasoning models available after the AIMO deadline in March 2026. The participation in FMI will be open to everyone interested in contributing to FMI's mission as described in a 1–2-page technical proposal submitted to us by the participants.[11] Proposals from the participants will be assessed by their contribution to the FMI's mission. Given enough resources, we will provide a justified resource allocation to every relevant proposal. We will provide all acceptance decisions within a week from the proposal's submission date.

3 The FMI Interpretability Challenge

The Fields Model Initiative provides a unique opportunity to open Interp research to the broader AI community beyond the best-resourced labs by providing everyone with resources necessary for Interp research relevant in 2026. The researchers participating in FMI will be provided with:

1. Generous compute resources of 128 H200 GPUs (or more, if justified in the proposal) allowing everyone to analyse and interpret the most relevant, billions-parameter-scale models;
2. An unrestricted access to containerized, easily reproducible AI systems, including the weights of SoTA reasoning LLMs submitted to AIMO;
3. Training reports detailing the approaches of winning teams and a direct communication channel with the development teams within FMI;
4. Access to intermediate and final model checkpoints from FMI's reproduction studies and ablations (data, architectures and/or training objectives).

This will empower everyone in the AI community to set off for much more ambitious and previously impossible goals. These may involve methods such as model diffing and convergence analysis at scale, behavioral and mechanistic analyses,

[11] Proposals will be collected through the FMI website: https://www.fieldsmodel.org

ablations of the emergence of models' capabilities in response to concrete refinements in training strategy, and more. The involvement and support of Interp research under the FMI aligns well with FMI's mission; to deepen our understanding of frontier reasoning models, allowing us to draw new insights into their robustness and covariates of their success.

4 The FMI Eval and Interp Competition

The FMI Eval and Interp Competition, organized within FMI, is our response to the need for quantitative grounding in Interp research. The goal of the challenge is to quantify, compare, and incentivise the development of Interp methods in drawing *actionable* insights; in our case, through their ability to uncover the practical robustness of LLMs.

Task formulation Given a mathematical problem and two models (incl. their weights) both providing a correct solution to this problem, the task is to **decide which of these two models responds to the given problem *robustly*.**

To make this possible, we will first train analogical versions of the same LLM in two variants; the *robust* and the *spurious* one:

1. reproducing the recipe of one of best-performing teams from AIMO (*Robust* model);
2. using the same recipe but with a substantial portion of data solvable by modeling a one of spurious correlations from a pre-defined set (*Spurious* model).

We will verify that the Spurious model indeed relies on a spurious feature exploitable for a given sample by *counterfactual assessment* – collecting mathematical problems with minimal adjustment that will (1) allow, and (2) disallow solving the problem using the treated spurious heuristic. The Competition's evaluation will only include problems for which we counterfactually verified that a Spurious model indeed relies on one of the spurious features from the collection.

The competing Interp methods will be evaluated by their *accuracy* to identify which of the models from the given pair is the *robust* one, i.e. the one that can *not* be misled by the counterfactual problem. The final ranking of different methods in the competition will be constructed from this accuracy. To make the challenge as accessible as possible, both variants of the models will maintain the same model types, so that the models internals can be assessed using an identical programming logic.

At the beginning of the competition, we will provide participants with a **validation** set of problems, each problem coupled with the pair of models and their per-token activations. After the competition, we will evaluate the submissions on a **test** set. The set of spurious features applied in the creation of Spurious models in the validation and test set will not be public a priori and will differ between the validation and the est set.

Using the identical evaluation framework, The FMI Eval and Interp Competition will distinguish two submission tracks: **black-box** and **white-box**.

1. The black-box track will be dedicated to behavioural methods – allowing the use only of the models' outputs, including per-token output probability distributions. This track will allow the use of any external data, including the labelled data.
2. The white-box track will additionally allow the submissions to utilize models' weights and activations. However, the submissions will not be allowed to use any labelled data beyond the given problem. Using the unsupervised data will be allowed (e.g. for fitting sparse autoencoders or transcoders for each of the models in the pair).

The FMI Eval and Interp Competition will be organized as a shared task within FMI, but participation in the Competition is independent from the participation in FMI. Proposals for participation in FMI will also be judged independently of the participants' interest to participate in the Competition.

5 Conclusion

Fields Model Platform opens up a unique opportunity to harvest and realize the ideas of the wider AI community in evaluation and interpretability that bear the potential to accelerate research towards better reasoning AI. We invite both researchers and practitioners that share our mission – to understand and improve the reasoning of LLMs and beyond – to embrace this opportunity, to put together and submit a short proposal to FMI at https://www.fieldsmodel.org[12].

References

1. Anand, Cappellazzo, U., Petridis, S., Pantic, M.: Mitigating attention sinks and massive activations in audio-visual speech recognition with llms (2025), https://arxiv.org/abs/2510.22603
2. Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=78Nn4QJTEN
3. Jiang, Z., Gu, J., Pan, D.Z.: Normsoftmax: Normalize the input of softmax to accelerate and stabilize training (2023), https://openreview.net/forum?id=4g7nCbpjNwd
4. Kadlčík, M., Štefánik, M., Mickus, T., Kuchař, J., Spiegel, M.: Pre-trained language models learn remarkably accurate representations of numbers. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 26693–26702. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1356, https://aclanthology.org/2025.emnlp-main.1356/

[12] The submission form will be made available on the FMI website

5. Mickus, T., Vazquez, R., Attieh, J.: I have an attention bridge to sell you: Generalization capabilities of modular translation architectures. In: Tafreshi, S., Akula, A., Sedoc, J., Drozd, A., Rogers, A., Rumshisky, A. (eds.) Proceedings of the Fifth Workshop on Insights from Negative Results in NLP. pp. 34–40. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.insights-1.5, https://aclanthology.org/2024.insights-1.5/
6. Ruscio, V., Nanni, U., Silvestri, F.: What are you sinking? a geometric approach on attention sink. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=OiC78C68sJ
7. Wendler, C., Veselovsky, V., Monea, G., West, R.: Do llamas work in English? on the latent language of multilingual transformers. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15366–15394. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.820, https://aclanthology.org/2024.acl-long.820/
8. Štefánik, M., Mickus, T., Kadlčík, M., Højer, B., Spiegel, M., Vázquez, R., Sinha, A., Kuchař, J., Mondorf, P.: Unravelling the mechanisms of manipulating numbers in language models (2025), https://arxiv.org/abs/2510.26285